Goldstein, H. (2015). Ofqual's reliability compendium. *Assessment in Education: Principles, Policy and Practice*, *22*(4), 495-497. https://doi.org/10.1080/0969594X.2014.970613

Review for AIE of 'Ofqual's reliability compendium'

Edited by Dennis Opposs and Qingping He.

Available for reading: http://ofqual.gov.uk/standards/research/reliability/compendium/

Reviewer:

Harvey Goldstein

University of Bristol

GSOE

35 Berkeley Square

Bristol, BS8 1JA

h.goldstein@bristol.ac.uk

This 900 page volume contains 22 contributions on the theme of reliability in educational assessment. OFQUAL, the English statuary body responsible for oversight of educational examinations and national test programmes, ran a programme from 2008-2011 to review the ways in which 'reliability' was defined and used. The contributions cover OFQUAL's own programme, including public examinations and key stage tests, technical articles on the definition and estimation of reliability measures and public perceptions of reliability. In a final chapter the editors attempt to summarise the contributions.

As its title suggests the book is not a text for people who wish to take a course in understanding and using reliability measures. It is, for a start, very repetitive, discursive and not designed to guide readers through it in a structured fashion. The first two chapters are meant to be scene setting. They attempt to present, in non-technical language, various different concepts of 'reliability', mainly in the form of summary reports of various working group discussions. They are written, it seems, by a variety of people and as such really don't present a clear picture of the terrain. In fact, in places they only succeed in causing confusion as in the first chapter where '(less than perfect) comparability of examinations from one year to the next' seems to be regarded as an example of reliability – which is hardly helpful.

Chapter 3 discusses different ways to assess reliability of Key Stage 2 tests (for 11 year olds) in England. Interest in this is somewhat specialised and it does not attempt to draw any general conclusions. Chapter 4 is similar, but focussing on the probabilities of the test scores misclassifying students into key stage levels. The next chapter applies similar procedures to an analysis of public examination looking simply at the characteristics of the exams themselves, ignoring other factors such as marker reliability. The following chapter generalises somewhat by looking at marker as well as test based factors. Unlike most of the other contributions in this volume this chapter provides a very clear definition of its concepts by setting out the underlying statistical models used. Unfortunately the choice here of so-called 'generalisability theory' is unfortunate since it is thirty years out of date and statistically unsound. Modelling the components of variability is just a special case of general random effects (multilevel) models and there are plenty of software packages that can do this properly.

Chapter 7 is a summary of various analyses – again of very specialist interest. It contains some very basic definitions (e.g. of Cronbach's alpha) but, as with much of the literature in this area, uncritically. For example the notion of 'parallel' tests is introduced, without any definition, and then it is assumed that such things actually exist so that some simple formulae can be used to provide reliability estimates. The second part of the chapter goes on, briefly, to estimate reliability by using Rasch models, justifying this rather curiously by the fact that the Rasch score estimates are a

monotonic function of the total test score – and unsurprisingly obtaining rather similar results. Perhaps the most interesting part of this chapter is the presentation of marker differences.

Chapter 8 looks at reliability of vocational assessment. Here the evidence is on discrete decisions for example mastery/non-mastery. An interesting study is described looking at assessor agreement, with an interesting discussion of practicalities. The next chapter looks at the reliability of teacher assessment in public examinations. This is a thoughtful chapter, with quite general implications for judgementally based assessments.

The next section of the volume starts with a chapter setting out some of the traditional models for reliability definition and estimation. This will be useful as a basic introduction, although there are numerous texts on measurement that contain the same material. The following chapter is somewhat similar and re-presents much of the earlier (unhelpful) material on generalizability theory. In similar vein is the next chapter on composite score reliability.

The following section of the volume reports the results of a series of surveys into how reliability measures are reported. There are some interesting things that emerge here. Rather depressingly it is reported (as many would suspect) that few attempts are made to present reliability caveats attached to test scores. There are some interesting examples of reporting for licensing exams in the US where some examples of good practice exist.

There are some interesting chapters on public communication of reliability concepts with the suggestion that the public is able to understand quite sophisticated concepts, but rarely given encouragement so to do. The public in general appear not to know how much unreliability there is in exam grades, although people appear fairly realistic in accepting that some 'random' error is inevitable. Very similar issues emerge in, for example, public understanding of things such as school league tables, and one suspects, as is hinted at here, that short term political considerations usually triumph over concerns to provide really useful information. There seems to be a considerable need here for public education. This section of the volume is really the most useful and OFQUAL might well consider revising it so that it tells a coherent narrative and issuing it as a report in its own right aimed at a programme of action.

There is a useful chapter setting out the final report of OFQUAL's technical advisory group. This looks at what could be done to improve practice and makes several recommendations and generally is in favour of more transparency. A chapter by the policy advisory group generally supports transparency and seeks to encourage continuing research. There is a final chapter by the editors that seeks to summarise the report.

Overall, this volume does contain much of interest amid a great deal of repetitive and often fairly trivial material. It does make some worthy suggestions about presenting reliabilities, although most of these are fairly obvious.

This compendium, and the report that underpins it, could have been much more. Thus, given the ubiquity of less than perfectly reliable assessments, how should this affect what selectors, trainers and students themselves do? Does an exam grade with a potentially large amount of measurement error (leaving aside considerations of cultural bias etc.) deserve to be down-weighted in a future selection procedure? Can a student with a 'poor' grade on an unreliable assessment justifiably appeal a decision on their future on this basis? Should we be prepared to trade off potentially less reliable teacher judgements against more reliable 'objective' tests for the sake of more authenticity and potential corrections for 'unfairness'. These, and similar, are questions that have always been there, even if carefully avoided because of the difficult issues they raise. It is a pity that OFQUAL did not take the opportunity to discuss such issues. It is a pity, but, as it stands, it seems unlikely that the present compendium will do much to promote such a debate.